The Social Mechanisms of Trust

Barrera, Davide

Published in:
Sociologica

DOI:
10.2383/27728

Focus

The Social Mechanisms of Trust

by Davide Barrera

doi: 10.2383/27728

1. Introduction

In the 1990s, when I was an undergraduate student, in front of the main building of the University of Turin there was a big parking lot, which was always completely full. Two guys were looking after the parking lot every day from early in the morning till very late in the afternoon. They were neither authorized, nor paid by the town council, but they received tips from the students to optimize the space in the parking lot and monitor their cars. The guys did not look very reliable, they wore poor and dirty clothes and they were covered by self-made aggressive-looking tattoos. However, it was not unusual to see one of these guys sitting in a very expensive sports car, smoking stinky cigarettes and listening to loud music from the car-stereo. In fact, when the parking lot was completely full, students used to leave their car to the guys who parked it as soon as a place was available. Imagine a first-year student, who goes to the university by car for the first time, enters the parking-lot and finds out that there is no place available to park her car. Then, one of the unauthorized "attendants" offers her to take the car, park it as soon as a place is available, and then leave the key on it. Judging from the appearance, the guy certainly does not look trustworthy. However, it seems that many other students indeed trust him, since there is a long queue of cars waiting to be parked, including some expensive ones. Would the student leave her car to the attendant? Would she take the number of other cars waiting to be parked as a signal that the attendant is trustworthy?

This example represents a typical *trust problem* between two actors. By "trust problem" I indicate a situation with a specific incentive structure which I will specify later. Moreover, the trust problem described in this example occurs in an *embedded* setting. I refer to this situation as "embedded" because it does not represent a simple isolated encounter between two actors. On the contrary, in the example, the actors may know each other (e.g., because the same interaction occurs every day), and they may share relations to some third parties (e.g., the other students who left their cars to the parking-lot attendants). Consequently, some information concerning the actor to be trusted or the specific trust problem is available to the actor who has to decide whether to trust.

2. Trust

In recent years, research programs on trust have been extremely numerous and diverse in terms of both theoretical and methodological approaches, and empirical applications. Moreover, research on trust has spread widely across disciplines [Barrera 2005; Bijlsma and Costa 2005; Bijlsma and van de Bunt 2003; Bohnet and Huck 2004; Bolton *et al.* 2004; Burt 2005; Burt and Knez 1995; Buskens and Raub 2002; Das and Teng 1998; Gulati 1995; Macy and Skvoretz 1998; Möllering 2005; Nooteboom 2002; Rus and Iglič 2005; Simpson and McGrimmon 2007; Snijders and Keren 2001; van de Bunt *et al.* 2005; Wittek 2001; Yamagishi and Yamagishi 1994]. Overviews can be found in Misztal [1996], Rousseau *et al.* [1998], and Buskens and Raub [2008]. In this paper, I will focus on one research stream and advocate approaching the problem of trust from the point of view of the individual actors involved, in order to gain a better comprehension of the process by means of which trust processes develop, stabilize, and collapse.

The approach I advocate is generally indicated by the label *Analytical Sociology* [Hedström and Swedberg 1998; Barbera 2004; Hedström 2005]. In the framework of this approach, what qualifies a sociological explanation is a focus on collective phenomena that result, often as unintended consequences, from the actions of the individuals who are restricted by the constrains and opportunities imposed by the social system in which the collective phenomenon emerges [cf. Boudon 1986; Coleman 1990; Hedström 2005]. The structure of an analytical explanation implies that three sets of assumptions need to be made explicit in order for a collective phenomenon to be understood: 1) a (micro) theory of action, specifying the principles regulating individual actions or decisions; 2) a macro-to-micro transition, defining how individual actions are restricted or influenced by the environment in which they are embed-

ded; 3) a micro-to-macro transition, an aggregation rule that determines how a set of individual actions combine to produce a collective outcome [Coleman 1990, ch. 1]. The combination of these three sets of assumptions identifies the *Social Mechanisms* determining the emergence of the collective phenomenon [Hedström and Swedberg 1998; Barbera 2004; Hedström 2005].

In the remainder of the paper, I will begin by introducing the key ingredients that are necessary to spell out the social mechanisms of trust. Then, I will summarize the mechanisms operating in trust processes. Subsequently, I will briefly review the empirical research on these mechanisms and finally I will conclude with a discussion on the usefulness of this approach.

2.1. *What is Trust?*

In the literature on trust, a first distinction could be made between scholars focusing on the function of trust for the social system and those that look at trust at the individual level. Examples of the first type of approach are Parsons' conception of trust in the normative system as a source of social order [e.g. Parsons 1937] and Luhmann's argument that trust serves the purpose of reducing complexity which characterizes modern societies [Luhmann 1988; see Misztal 1996, ch.3, on the different functions of trust]. However, as I anticipated in the introduction, social mechanisms cannot be understood unless one identifies the relevant units of analysis: the individual actors and the social interactions in which they are involved [Hedström 2005]. As far as the study of trust is concerned, actors can be individuals or collectives (e.g., corporates, local institutions, national states, etc.). A problem of trust is a social interaction involving at least two actors: a *trustor* and a *trustee* [e.g., see Snijders 1996; Buskens 2002; Buskens and Raub 2002; Barrera 2005]. I will further clarify the differences between these two roles and the properties of the social interaction in which they are involved in the remainder of this section.

Among scholars who studied trust at the micro level, one can find, on the one hand definitions of trust that focus on psychological and cognitive elements [e.g., Barber 1983; Lewis and Weigert 1985; Robinson 1996] and, on the other hand, definitions that stress strategic and calculative elements [e.g., Arrow 1974; Camerer and Weigelt 1988; Dasgupta 1988; Kreps 1990; Williamson 1993]. All of these definitions more or less uniformly identify an element of risk due to the fact that, as far as the trust problem is concerned, the welfare of the trustor *depends* on the action of the trustee [cf. Rousseau *et al.* 1998, 395]. However, insofar as they fail to take the incentives of the actors involved into account, most definitions that treat trust as a psychological

state apply also to situations that, in my view, are not trust problems. For example, Robinson [1996, 576] defines trust as a person's "expectations, assumptions, or beliefs about the likelihood that another's future action will be beneficial, favorable, or at least not detrimental to one's interests." Then, according to this definition, we can say that, when we lie in a hospital bed, we trust the doctors that their actions "will be beneficial, favorable, or at least not detrimental to our interests." However, if we are considering the doctors' capability to make the right diagnosis and choose appropriate treatments, we should rather say that we *confide* that their actions will be beneficial, and not that we *trust* them, because – as stated by Snijders [1996, 10] using a similar example – we rely on their competence, not on their preferences. Conversely, if we are considering the possibility that doctors do not apply perfect care-intensity to our problem, for example because they want to dedicate more of their time to other cases, it is appropriate to say that we trust them, because the source of our concern lies in their preferences.

In the literature on trust, another important distinction can be found between scholars who treat trust as explanans and others who treat it as explanandum [Craswell 1993]. In the first case, trust is offered as an explanation for certain behavior, typically seemingly non-calculative behavior, such as cooperation in one-shot Prisoner's Dilemma. Trust is used as an explanans when it is invoked as an alternative justification for the trustor's risk-taking behavior, namely a justification that excludes the possibility that such risk is in the trustor's calculated interest [e.g. Lewis and Weigert 1985]. Trust is typically conceived as an explanans by scholars who study trust at the level of the system [e.g. Parsons 1937; Luhmann 1988]. Conversely, in the second case, the term trust indicates risk-taking behavior in situations that are described as a subclass of risky situations, namely those in which the risk to which the trustor exposes himself depends on the performance of another actor [Coleman 1990, ch. 5]. However, when trust is seen as an explanandum, the term trust is used only to *describe* the trustor's risk-taking behavior, rather than to offer an *explanation* for it. The explanation of the trustor's trusting behavior requires that some theory is provided, which may very well include a calculative explanation [e.g. Gambetta 1988; Coleman 1990]. In Craswell's [1993] terms, this paper deals with trust as an explanandum. Understanding the social mechanisms of trust means explaining the trusting decision of the trustor (explanandum) as a function of the information about the trustee and about

the social interaction that is available to the trustor before her trusting choice is made.[1]

Gambetta [1988, 217] defines trust as "a particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action [an action that is beneficial or at least not detrimental], both *before* he can monitor such action, [...] *and* in a context in which it affects *his own* action" (emphasis in the original). Although this definition still includes both "trust in the intentions of others not to cheat us and in their knowledge and skill to perform adequately over and above their intentions" [*ibidem*, 218], Gambetta adds two important elements. First, he restricts trust to situations in which the *action* of the trustor depends on such subjective probability, and second, he introduces time asymmetry by specifying that such subjective probability is assessed *before* the action of the trustee can be monitored. In other words, trust is conceptualized as a strategic decision of the trustor based on the subjective probability that a certain event – dependent on the will of the trustee – will occur, and not just as a psychological state.

Although his definition is merely cognitive and trust as behavior is explicitly excluded, Hardin adds another important element in his definition of trust: the trustworthiness of the trustee. If the trustor's trust in the trustee depends on the trustor's assessment of the trustee's trustworthiness, then it is essential to understand trustworthiness in order to explain trust. Thus, a congruent definition of trust must take the trustee's incentive to be trustworthy into account. According to Hardin [2002, 4], trust is *encapsulated interest*: "I trust you because I think [... that] you have an interest in attending to *my* interest because, typically, you want our relation to continue." Then, Hardin explores a number of possible arguments supporting the trustee's trustworthiness. These arguments include internal motivations, such as dispositions, moral rules and internalized norms, as well as external motivations, such as institutional devices and commitments [cf. Raub 2004].

The major features of a trust problem are summarized by Coleman [1990, ch. 5], who identifies four essential characteristics:

-The trustor has the possibility to place some resources at the disposal of the trustee who has the possibility to honor or abuse trust.

-The trustor prefers to place trust if the trustee honors trust, but regrets placing trust if the trustee abuses it.

[1] For reader friendliness, throughout this paper I will refer to the focal actor facing the trust problem (the trustor) using female personal pronouns, and to her partner, the actor who is to be trusted (the trustee), using male personal pronouns.

-There is no binding agreement that protects the trustor from the possibility that the trustee abuses trust.

-There is a time lag between the decision of the trustor and that of the trustee.

The first point stresses the dependency of the trustor's welfare on the action of the trustee. The second point specifies that a trust problem is characterized by "mixed motives", in the sense that the interests of the trustor and the trustee are partly common and partly conflicting. Consequently, a trust problem implies the risk that the action of the trustee harms the interests of the trustor. The third point emphasizes that this element of risk cannot be eliminated exogenously, and the fourth point restricts the definition to situations with a sequential structure, excluding situations in which the decisions of the trustor and the trustee take place simultaneously. It follows from this definition that the trustor's trust in the trustee is the trustor's *decision* to place some resources at the disposal of the trustee when confronted with a situation resembling the description above. According to Coleman this decision depends on the trustor's subjective probability that the trustee will honor trust and on the possible gains and losses depending on the trustee's decision. The subjective probability that the trustee will honor trust represents the trustor's assessment of the trustee's trustworthiness. Coleman discusses possible reasons for the trustee to be trustworthy throughout the chapter. However, he does not model explicitly the incentives of the trustee. Formal definitions of a trust problem incorporating also this aspect can be found in game-theoretical models.

2.2. *Models of a Trust Problem*

Game Theory is a very powerful instrument to investigate micro processes and social interactions between interdependent actors. More specifically, the "games" constitute a taxonomy of stylized social interactions, while the "theory" consists of a set of rules and tools generally used to describe or prescribe how actors should behave when playing the games under various assumptions concerning their rationality and preferences [Camerer 2003]. The simplest game-theoretical models assume 1) that actors are *rational*, in the sense that their preferences can be consistently rank-ordered and 2) that actors are *selfish*, in the sense that they are not interested in the payoffs obtained by the other [Fehr and Gintis 2007; Buskens and Raub 2008; Gächter 2008]. In general the rationality assumptions include also the assumption that actors are *forward-looking*. This means that their decisions are only guided by the expectations concerning future benefits while they take information about the past into account only as far as it helps computing expected benefits. Furthermore, many

game-theoretical models also assume that actors possess all the relevant *information* concerning own and other's preferences and set of alternative actions in a given game (i.e., common knowledge and perfect information). This set of assumptions constitutes the *decision theory*. In any social interaction – modeled using a particular game –, the combination of the (predicted) decisions of the actors involved constitutes the *solution* of the game. The solution of a game is an *equilibrium* when, given the decision of all the other actors involved, no agent has an incentive to change her decision [e.g., Gächter 2008].

Standard game-theoretical models are generally quite parsimonious and allow for precise predictions for most games, but they do not describe the actual behavior of real actors very well. They are the social sciences equivalent of the perfect gas model in chemistry: they provide a useful analytical tool and a normative benchmark.[2] In order to improve the fit between theoretical models and empirical observations, various solutions have been proposed: some scholars have argued that the actors' rationality is actually limited and they have explained the empirical anomalies in terms of the actors' failure to recognize the incentive structure of the game being played [e.g. Binmore 1998]; Other scholars have opted for a modification of the second assumption, and have proposed alternative models of actors with somewhat altruistic preferences [e.g. Rabin 1993; Fehr and Schmidt 1999; Bolton and Ockenfels 2000]. This set of alternative models and assumptions, inductively modified by looking at the actual behavior of subjects in experiments, is commonly indicated as Behavioral Game Theory [Camerer 2003]. In general, when studying social interactions, one starts with the simpler models and then modifies the underlying assumptions stepwise in order to improve their predictive capability. The Games used to study trust problems are the *Trust Game* [Camerer and Weigelt 1988; Dasgupta 1988; Kreps 1990] and the *Investment Game* [Berg *et al.* 1995; Ortmann *et al.* 2000].

[2] I owe this metaphor to David Willer (personal communication).

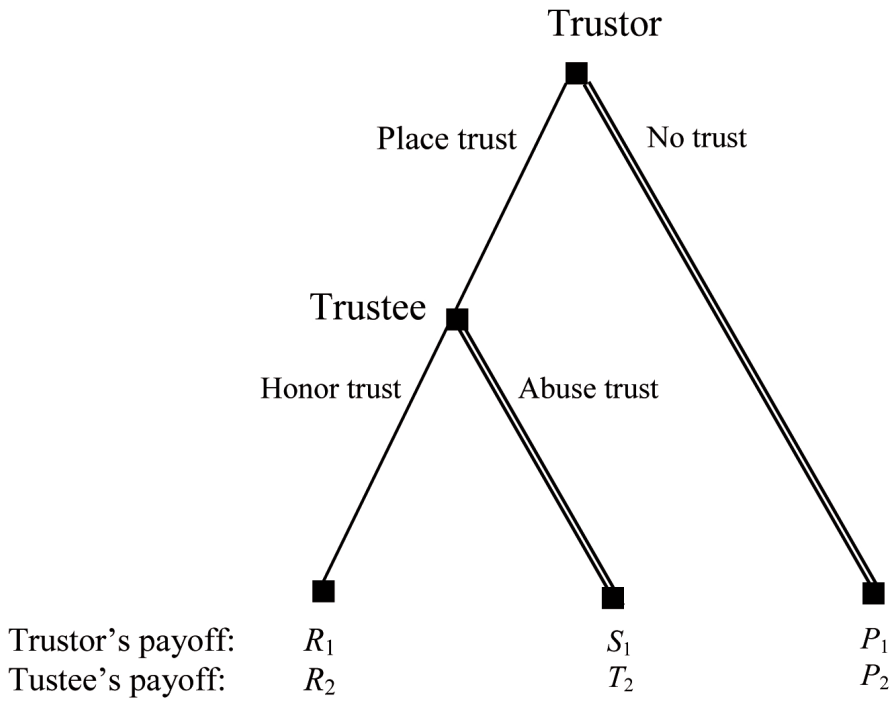

FIG. 1. Trust Game ($T_i > R_i > P_i > S_i$).

The Trust Game (Fig.1) begins with a move by the trustor who has a choice between trusting and not trusting the trustee. If the trustor withholds trust, the game ends. In this case, the trustor receives P_1 and the trustee receives P_2. If the trustor chooses to place trust, the trustee has the possibility to honor or abuse trust. If the trustee honors trust, he obtains

R2 > P2

and the trustor obtains

R1 > P1,

while if he abuses trust the trustee receives

T2 > R2

and the trustor is left with

S1 < P1.

This game can be seen as a one-sided version of the well-known Prisoner's Dilemma. For this reason, payoffs are indicated with the conventional letters used in the literature on the Prisoner's Dilemma: *T* for *temptation*, *R* for *reward, P* for *punishment,* and *S* for *sucker*. Generally speaking, the trusting decision of the trustor depends on her subjective probability that the trustee is trustworthy [Gambetta 1988; Coleman, 1990, ch. 5]. The assessment of this probability is based on what the trustor

knows about the trustee's preferences [Hardin 2002] and about the structure of the strategic interaction. If the game is played only once (i.e., there is no common past and no common future) between two isolated actors (i.e., there are not any third parties involved in the social interaction), and payoffs equal utilities, on the basis of the standard assumptions listed above, the trustor has no reason to expect the trustee to be trustworthy.[3] Then, it is predicted that the trustor should not place trust: if the trustor placed trust, the trustee would in fact abuse it because

T2 > R2.

Consequently, the trustor – knowing the payoff structure – should withhold trust because

P1 > S1.

"No trust" and "Abuse trust" are equilibrium choices (in Fig.1 this is represented by double lines). Therefore, the payoffs in equilibrium are P_1 and P_2. This outcome is Pareto sub-optimal, because both actors would prefer the payoffs yielded in the situation in which trust is placed and honored, R_1 and R_2.

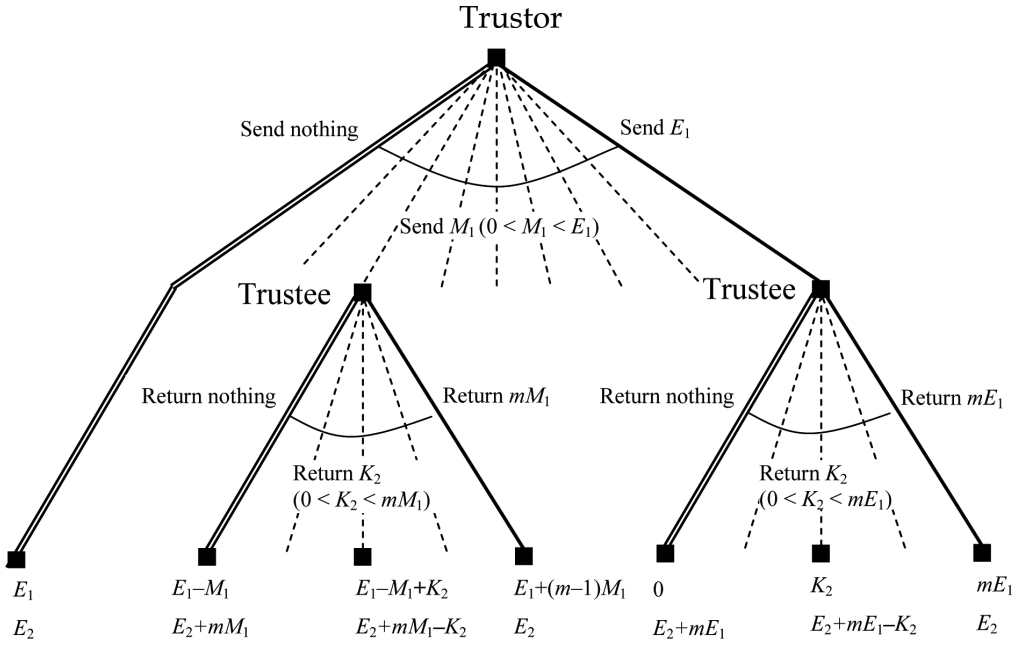

FIG. 2. Investment Game.

[3] Assuming that payoffs are utilities implies that any sort of moral, emotional, or psychological concern – such as envy, guilt, regret, fairness, etc. – induced by the outcomes of the game, does not alter the rank order of the payoffs or, in other terms, all possible moral, emotional, or psychological concerns are already incorporated in the payoffs so that the preference order of the players is T > R > P >S. For examples of theoretical models of Trust Games explicitly incorporating such psychological aspects see Snijders [1996].

The Investment Game (Fig.2) models a situation in which the trustor's choice whether to trust the trustee and the trustee's choice whether to honor trust are not dichotomous choices as in the Trust Game. The two players start with an initial endowment, E_1 and E_2. The trustor has then the possibility to send all, some, or none of her endowment to the trustee. The amount of money that she decides to send, denoted M_1, is then multiplied by a factor m (with m > 1). The trustee receives an amount equal to m times the amount sent by the trustor. The parameter m can be interpreted as the returns the trustee makes due to the investment of the trustor. Then the trustee can decide to send back to the trustor all, some, or none of the money he has received. The amount returned by the trustee – denoted K_2, satisfying

$0 \leq K_2 \leq mM_1$

– is not multiplied. After players have chosen their actions, the game ends and the payoffs are computed.[4] The payoff earned by the trustor (V_1) is:

V1 = E1 M1 + K2,

whereas the payoff earned by the trustee (V_2) is:

V2 = E2 + mM1 K2.

The amount that the trustor is willing to send to the trustee indicates the extent to which the trustor trusts the trustee. Therefore, I refer to the trustor's choice M_1 as (degree of) *trust*. Conversely, the amount that the trustee is willing to return to the trustor represents the extent to which the trustee is trustworthy. Therefore I refer to the trustee's choice K_2 as (degree of) *trustworthiness*.[5]

As in the trust game, assuming a one-shot game in which payoffs equal utility, the trustee maximizes his revenues by keeping everything the trustor has sent to him. Thus, the trustee should choose

K2 = 0.

Consequently, knowing the structure of the game and anticipating the trustee's behavior, the trustor maximizes her revenues by choosing

M1 = 0, since

E1 M1 < E1 if M1 > 0.

Therefore, "Send nothing" and "Return nothing" are the equilibrium choices (in Fig. 2 this is represented by double lines) and the payoffs in equilibrium are E_1 and

[4] In figure 2, the payoffs earned by the trustor and the trustee are displayed above each other, next to the end nodes of the game.

[5] In the economic literature, K_2 is often labeled *reciprocity* [e.g. Berg *et al.* 1995; Ortmann *et al.* 2000]. The term reciprocity is used because if the trustor chooses a small M_1, the trustee might choose a small K_2 as well in order to punish the trustor for not trusting him. Thus a small K_2 does not necessarily mean that the trustee is not trustworthy. However, the term reciprocity implies some psychological speculation about the cause of the trustee's choice K_2. Therefore, I prefer the term *trustworthiness*.

E_2. As in the Trust Game, this outcome is Pareto-suboptimal, because both actors would prefer the payoffs yielded in the situation in which trust is placed and honored,

E1 M1 + K2

and

E2 + mM1 K2,

with

M1 > 0

and

K2 > M1.

For the Investment Game, Pareto improvements are always possible if

M1 < E1.

The pie that the actors are dividing reaches its maximum when the trustor sends everything

(M1 = E1).

Ego gains from trusting the trustee only if the trustee returns more than what the trustor sent

(K2 > M1) ,

but, given

M1 = E1,

all possible K_2 chosen by the trustee induce outcomes that are Pareto non-comparable.

These two games differ because in the Trust Game "trust" and "trustworthiness" are represented by dichotomous choices – trust vs. no trust, honor trust vs. abuse trust –, while the Investment Game exhibits some "continuity" both in the choice of placing trust and in the choice of honoring or abusing trust. The (symmetric) Trust Game presented above can be seen as a "special case" of the Investment Game in which the trustor has to decide whether to send everything or nothing to the trustee,

M1 = E1 or M1 = 0,

and the trustee – if the trustor chooses to send everything – can choose to keep everything

(K2 = 0),

or split the amount received in such a way that both actors end the game earning the same payoff

(K2 = ½mE1 + ½E2).[6]

[6] Asymmetric situations in which the trustor and the trustee do not earn the same when trust is placed and honored ($R_1 \neq R_2$) can also be represented using the Trust Game (e.g., see Bohnet and Huck

Both games represent social dilemmas characterized by a conflict between individual and collective rationality [Rapoport 1974]. The situation in which trust is placed and honored is collectively rational because in such a situation both actors obtain a better payoff in comparison with the situation in which trust is not placed. Nevertheless, it is individually rational for the trustee to abuse trust if the trustor places trust. Consequently, it is individually rational for the trustor to withhold trust. Thus, if the actors are individually rational and expect their partner to be individually rational, a social dilemma yields a collectively irrational outcome.

Thus, whatever game one uses to model the trust problem, standard Game Theory predicts that the trustor does not place trust and, if she did, the trustee would abuse trust. Does this imply that according to standard game theory trust cannot emerge? No, it only implies that, according to standard game theory, trust cannot emerge in situations in which all the assumptions specified above are met. More specifically, it implies that trust cannot emerge in isolated encounters between perfect stranger who meet only once and have a certain preference order for the payoffs of the game. However, this is clearly a rather stylized situation. In real life, most trust problems occur in *embedded* settings, rather than in isolated encounters, the information the actors posses may be *incomplete and/or asymmetric* rather than complete and symmetric, the actors may take their *past* into account rather than being purely forward-looking, the actors' preferences may be partly *altruistic* rather than purely selfish.

3. Trust in Embedded Settings

3.1. *Types of Embeddedness*

Unlike economists and (social) psychologists, sociologists are primarily interested in collective phenomena, rather than in micro processes. Then, since standard Game-theoretical models of trust do not describe very accurately the behavior of the actors, the models should be improved by injecting more realism into the assumptions concerning the macro level of the explanation rather than into the assumptions concerning the micro level. Recall that analytical explanations include three components: 1) a set of rules regulating individual decisions (behavioral theory), 2) a set of assumptions concerning how the social conditions impact on the individual decisions (macro to micro transition), 3) a set of assumptions on how individual decisions ag-

[2004]). By contrast, asymmetric situations in which trust is placed and honored can occur in the laboratory when subjects play the Investment Game (if $M_1 > 0; K_2 > M_1; E_1 - M_1 + K_2 \neq E_2 + mM_1 - K_2$).

gregate (micro to macro transition). Therefore, when a simple model does not work, one should start improving it by modifying assumptions concerning the second and third component. By contrast, the behavioral theory can be kept simple, as much as possible.

For the study of trust, this means changing the assumptions on social conditions replacing social isolation with *social embeddedness* [Granovetter 1985; Raub 1997; Raub and Weesie 2000], rather than (or before) starting directly to modify the assumptions concerning the rationality or the preferences of the actors.

Two dimensions of embeddedness can be distinguished: first, the trustor and the trustee can have repeated interactions with each other, and, second, the trustor can have some relations with other actors. I refer to the first dimension as *dyadic* embeddedness and to the second as *network* embeddedness [Raub 1997; Buskens 2002; Buskens and Raub 2002; 2008; Barrera 2005]. Dyadic embeddedness refers to situations in which a relation between the trustor and the trustee pre-exists the specific trust problem, or to situations in which the trustor and the trustee are likely to be facing each other again after the specific trust problem will be solved. Conversely, network embeddedness refers to situations in which there exists at least one third party who is connected to the trustor by means of a relationship allowing him or her to provide the trustor with information about the trustee, as well as to receive similar information about the trustee from the trustor. In the next section I will illustrate the most important social mechanisms affecting the emergence of trust in embedded settings. First, I will discuss the mechanisms operating under conditions of dyadic embeddedness and then under conditions of network embeddedness. Introducing the various mechanisms, I will modify the baseline assumptions in a stepwise fashion. While discussing the social mechanisms I will generally refer to the Trust Game as a model of the social interaction. However, the mechanisms that I will discuss can be applied to the Investment Game, likewise.

3.2. *Dyadic Embeddedness and Social Mechanisms*

3.2.1. Dyadic Control

Dyadic embeddedness implies that the interaction between trustor and trustee is repeated rather than one-shot. An important distinction is whether the game is repeated a finite or an indefinite number of times. I discuss the indefinitely repeated case first and the finitely repeated case in the next paragraph. Repetition changes the analysis of a game considerably, because if a game is repeated, the actors need to take into account the consequences that their actions can have for the future stages of the

game. In a repeated game, the actors can make their action conditional on the behavior of their partner (i.e., technically, they can use conditional strategies). For example, a trustor may decide to place trust as long as the trustee honors it, but stop placing trust, as soon as the trustee abuses it. Thus, the trustor can exercise *control* over the trustee. Control implies that a trustor rewards trustworthiness in the present by placing trust in future games, and punishes abuse of trust in the present by withholding trust in future games. It has been shown that if the actors attach enough importance to future payoffs, the indefinitely repeated trust game has an equilibrium in which trust is always placed and always honored [Kreps, 1990; see Buskens and Raub 2008 for a detailed summary]. This equilibrium is not unique, but it pareto-dominates all other possible equilibria.[7] Thus, if an encounter is repeated, a control mechanism ensures that trust emerges. In terms of my introductory example, if the parking-lot attendant knows that I will be leaving my car under his custody for years to come, it is in his interest to take good care that my car is not stolen. Accordingly, since I know that he knows this, I can trust him. Of course, in this example, the control mechanism may not be sufficient to establish trust, because, assuming that he could simply steal the car himself, the parking-lot attendant would be weighting the value of my car today against a couple of Euros a day that I would give him as a tip until I graduate. His payoff corresponding to the temptation to abuse trust is much larger than the payoff for honoring trust.[8] However, such a simple control mechanism could be sufficient in situations where the temptation to abuse trust is not so high.

3.2.2. Dyadic Learning

While a mechanism of control can be sufficient to guarantee the emergence of trust in indefinitely repeated Trust Games, the situation changes drastically if one analyzes *finitely* repeated games. If a game is repeated a finite number of times but information is assumed to be complete, the game can be analyzed as a one-shot game because the argument of *backward induction* applies. Simply put, this means that, in a finitely repeated Trust Game if both actors have complete information (concerning preferences, alternative actions, outcomes, and duration of the game), the last stage of the game is identical to a one-shot game. But, if they know that there will be no trust in the last game, then the one but last game can likewise be analyzed as a one-shot game.

[7] The proof of existence of multiple equilibria in indefinitely repeated games is known as the Folk Theorem. A typical solution when multiple equilibria are present is to assume payoff dominance, which implies that both actors prefer to coordinate on the pareto-optimal equilibrium.

[8] It can be shown formally that if T – R is very high, "no trust" is likely to be the unique equilibrium of the indefinitely repeated Trust Game [Axelrod 1984; Kreps 1990; Buskens and Raub 2008].

Following this line of reasoning, the game unravels to the beginning and no trust can be expected already from the first game. Once again, however, standard game theoretical predictions do not describe the behavior of real agents very well. That is, real agents do not apply backward induction when playing finitely repeated trust games [e.g. Camerer and Weigelt 1988; Engle-Warnick and Slonim 2004]. However, behavior in such games is explained using a modified model in which information is assumed to be incomplete [Kreps and Wilson 1982; see Harsanyi 1967-1968 on the theory of games with incomplete information].

In a game with incomplete information, a move of *Nature* takes place before the players' moves and it is unobserved by at least one player [Rasmusen 2001, 50]. Typically, in games with incomplete information, it is assumed that different "types" of players exist. The move of nature consists in selecting which of the alternative types of players enters the game. In a Trust Game with incomplete information, it is assumed that there are two types of trustees, one of which has different alternative actions or different preferences concerning the payoffs ranking. For example, next to the standard trustee – who prefers T_2 over R_2 – there could be an alternative type for which

$T2 < R2$

or a trustee with no opportunity to abuse trust at all (see Raub [2004] for a model of a one-shot Trust Game with incomplete information). The trustor only knows the probability of encountering each type of trustee, but at the beginning of the game she does not know with what kind of trustee he is playing. This implies that the trustor is no longer sure that the trustee will abuse trust if trust is placed.

Kreps and Wilson [1982] developed a model for the finitely repeated Prisoner's Dilemma with incomplete information [see Bower *et al.* 1997 and Buskens 2003 for applications of similar models to a Trust Game]. Applying this model to a Trust Game, the trustee knows that if he abuses trust, the trustor *learns* that she is matched with a standard trustee – for which

$T2 > R2$

– and will stop placing trust in the next rounds. Conversely, the trustor keeps placing trust as long as the trustee honors it because she remains uncertain whether the trustee has an incentive to abuse trust at all. Thus, Kreps and Wilson's [1982] model predicts that, at the beginning of the finitely repeated game, trust is placed and honored for a number of rounds. In this phase all trustees honor trust. The ones without temptation honor trust because they cannot do otherwise, the standard trustees because they need to maintain a good reputation in order not to lose investments in later rounds of the game. When the end of the game approaches, reputation building becomes less important for the standard trustees, until they become indif-

ferent between honoring and abusing trust. At this point, all players start to randomize between their alternative actions. The randomization continues until trust is not placed or until it is abused and then no trust is placed until the end of the game. This solution, called *sequential equilibrium*, is a refinement of Nash equilibrium and holds also if the probability of encountering a non-standard trustee is very low.

I discussed the Kreps and Wilson [1982] model in this section on dyadic learning because the model assumes incomplete information and the possibility for the trustors to learn what type of trustee they are facing by observing the trustee's behavior in previous games. However, the standard game-theoretical rationality assumptions are still met in this model, i.e. actors are assumed to apply forward-looking rationality. By contrast, "pure" learning models typically assume backward-looking rationality [see Macy and Flache 1995 and Macy and Flache 2002 for applications of learning models to social dilemmas]. That is, they assume that actors do not look ahead and compute expected future payoffs, but they rather adjust their behavior according to their past experiences. Different types of learning mechanisms can be distinguished [see Camerer 2003, ch. 6, for an overview of such models]. The most widely applied families of learning models are *belief learning* and *reinforcement learning*. Belief learning models assume that actors update their belief about the other player's type or about the other player's expected behavior. Players then calculate the expected payoffs based on their beliefs concerning the other player's strategy and subsequently choose the strategy with the highest expected payoff. Conversely, reinforcement learning models are based on the payoffs that actors received in previous games: the higher the payoff obtained by a given decision, the more likely it is that a player will make that same decision again.

In a nutshell, one can say that the mechanism of dyadic control depends on the *shadow the future* [Axelrod 1984]: the length of the common future and the importance the players attach to it. Conversely, the mechanism of dyadic learning is based on the *shadow of the past*: the trustor is more likely to place trust if she has a history of successful interactions with a trustee, because she has learned that this trustee is reliable.

3.3. *Network Embeddedness and Social Mechanisms*

3.3.1. Network Control

Moving from dyadic to network embeddedness changes the scenario quite drastically: trustors can punish abuses of trust either by searching for another trustee or by informing other trustors about the behavior of the trustee. In Hirschman's [1970]

terminology, the former strategy corresponds to *exit* and the latter to *voice*. The availability of these options provides the trustors with a means to *control* the behavior of the trustees. The feasibility of the exit option depends on the availability of alternative partners and on the magnitude of sunk costs (i.e., relation-specific costs already incurred by the trustor). Modeling exit and voice simultaneously would require rather complex models. Moreover, in some situations, like the parking-lot example, the actors can only use voice because no alternative trustee is available. Therefore, I concentrate my discussion on models of voice [examples of models including exit can be found in Lahno 1995 and Weesie 1996].

As I anticipated in the beginning of section 3, my discussion proceeds stepwise: I began with standard game-theoretical assumptions and isolated one-shot games, and then I discussed repeated interactions without modifying the behavioral theory. In models with network embeddedness, the interaction is likewise repeated and one trustee interacts with more trustors [Buskens and Weesie 2000; see also Buskens, 2002, ch. 3] developed such a model for an indefinitely repeated Trust Game with a network of trustors. These models are variants of the reputation model proposed by Raub and Weesie (1990). As in the case of dyadic control, in this model the trustors use conditional strategies. However, the trustors' decision to place trust does not depend only on whether the trustee has previously abused trust *with her*, but also on whether the trustee has previously abused trust *with other trustors*. The model assumes in fact that the trustors are embedded in a network of relationships through which they can exchange information about the behavior of the trustee. This game-theoretical model predicts control effects via network embeddedness. The solution of the model implies that a given trustor places trust if the gain that the trustee would obtain by abusing trust is compensated by the losses he incurs due to the sanctions he will receive from the other trustors. In other words, placing trust is more likely if the sanction potential of a trustor is higher. The sanction potential depends on the extent to which that trustor transmits information to the network. Therefore, this model leads to testable hypotheses concerning the effects of properties of the network on the probability that a given actor places trust. For example, the probability that trust is placed increases with the density of the communication network in which the actor is embedded, and with the actor's outdegree in the same network.[9]

[9] In social network terminology, the outdegree of an actor is equal to the sum of her outgoing relationships. If the network considered is one of information transmission, the outdegree of an individual corresponds to the number of other actors to whom this individual transmits information. Conversely, density is a global property of the network and it corresponds to the extent to which all potential relationships in a given network are actually present or absent.

In another study, Buskens [2003] applied Kreps and Wilson's [1982] finitely repeated Prisoner's Dilemma model to a finitely repeated Trust Game. Buskens [2003] extended the original model by including an "exit" and a "voice" option for Ego. In the voice model, two trustors can inform each other about the trustee's behavior in previous interactions. This model assumes incomplete information just as in Kreps and Wilson [1982] and predicts that the trustor's decision to place trust increases with the frequency at which the two trustors can inform each other. Assuming that trustors have incomplete information – that is, there are two types of trustees, selfish ones and nice ones, the nice ones have no incentive to abuse trust – and that any abuse of trust is type-revealing, Buskens [2003] showed that the trustor's possibility to inform each other about the trustee's behavior makes the trustee more trustworthy than if the trustors played with a given trustee individually. Thus, while nice trustees do not abuse trust anyway, selfish ones mimic the behavior of nice trustees for longer than if they were only playing with one trustor, in order to keep a positive reputation.

In terms of the parking-lot example, these models imply that a student can more safely trust the parking-lot attendant the more she is embedded in social relationships with other students, i.e., the higher her sanction potential through voice. This conclusion applies under the assumption that the parking-lot assistant knows that the student is well embedded and thus can anticipate the he will receive high sanctions if he abuses trust. Although one can hardly argue that the rationality assumptions implied by these models are actually met, they present considerable improvements, in terms of empirical realism, compared with models of the one-shot situation. In fact, the solutions of these models permit that trust is placed and honored in equilibrium, even though the models still assume that the actors handle their decisions with standard game-theoretical rationality.

3.3.2. Network Learning

As I stated earlier, the trustor's decision to place trust in a given trust problem depends on her assessment of the probability that the trustee will honor trust [Hardin 2002]. In game-theoretical models this idea translates in the assumptions about the information that is available to the trustor before she makes her decision. Simple standard models of one-shot games do not leave any room for learning because the information is usually assumed to be complete. Conversely, learning is possible in models with incomplete information. If the interaction is repeated, the trustors can learn from own experience. If the trustors have relationships to other trustors with whom they exchange information about previous interactions, they can also learn from experiences made by others. It is perhaps reasonable to assume that such sec-

ond-hand information is less valuable than own experience, but it is certainly plausible that actors learn how to behave in certain circumstances by observing what others did and what payoffs did they obtain. The models typically used to study this mechanism are models of diffusion processes, e.g., diffusion of innovations [Valente 1995] and disease epidemics [Altmann 1993].

An application of this kind of model to the study of trust can be found in Buskens [2002, ch. 4]. Unlike the control models previously discussed, this model does not include any strategic decision. Instead, Buskens modeled the transmission of information focusing on the complexity of the social network through which the information travels. The model allows prediction about the speed at which a given actor receives a piece of information. Then, if one assumes that this information concerns the behavior of the trustee in previous transactions with other trustors, the model effectively predicts how fast a trustor learns as a function of the network structure in which she is embedded. Of course, the effects of the information received by a trustor on her trusting behavior depend only on the content of the information: positive information leads to more trust; negative information leads to less trust. The hypotheses – on the effects of network properties on trust – based on this model, mirror the hypotheses based on the control model. For example, trustors with a higher indegree are expected to be more likely to place trust (if the information about the trustee is positive) because they receive information faster.[10] One drawback of pure learning models is that since in these models trustors apply backward-looking rationality, the incentives of the trustee are not taken into account (a discussion of the link between learning and control models can be found in Buskens [2002, 102]). Conversely, learning models allow a more complex modeling of the network structure in which the actors are embedded.

For the parking-lot example, such a learning model implies that the student will be more likely to trust the attendant if she knows many other students who told her that they left their cars to the attendant and nothing bad ever happened to it.

3.3.3. Imitation

Most learning models include two important assumptions: 1) transmitting information is not costly and 2) the information that the actors receive is reliable. The costs of transmitting information need to be assumed away because otherwise the transmission of information would correspond to the production of a public good,

[10] In social network terminology, the indegree of an actor is equal to the sum of her incoming relationships. If the network considered is one of information transmission, the indegree of an individual corresponds to the number of other actors that transmit information to her.

implying a free-rider problem.[11] The second assumption implies that the potential incentives to disclose false information are neglected. How problematic are these two assumptions? They are certainly not problematic in those situations in which the behaviors of the trustor and the trustee are public or easily observable. For example, the reputation systems used by many online platforms make the reputation score of each seller easily available to all buyers (and possibly vice versa), the incentives to provide false information and the costs of providing the feedback are negligible.

However, in many instances, only the information about the behavior of other trustors is readily available, while the trustee's responses in those interactions are virtually impossible to obtain. When choosing where to have dinner in an unfamiliar city, one can easily observe which restaurant has more customers. However, one will never know how many of those customers are going to have a stomach ache the day after. It seems inappropriate to argue that trustors can actually *learn* about the trustworthiness of the trustees from such incomplete information. Nevertheless, it is perfectly plausible that this information on the behavior of other trustors leads to *imitation*, in the sense that an individual places trust in a trustee who is trusted by many others. Imitation is usually considered a form of learning that plays an important role in socialization processes [for example, Bandura and Walters 1963, ch. 2]. In interactions resembling social dilemmas, imitation could be viewed as a parsimonious way to achieve the optimal decision [cf. Hedström 1998 on "rational imitation"], especially in settings where information is scarce.

Barrera and Buskens [2007] proposed a distinction between learning and imitation bearing on the completeness of the information available. They adopt the term "imitation" to indicate situations in which available information does *not* include the outcomes obtained by others (e.g., other trustors facing a similar dilemma), but only their behavior. Conversely, they use the label "learning" for decisions based on "full" information that includes the outcomes obtained by others. Some imitation models have been proposed by economists [for example, Pingle 1995; Pingle and Day 1996; Schlag 1998], but they are actually variants of what Camerer [2003] would call "reinforcement learning" models because, in these models, actors make their decisions after receiving some information about the actions chosen by others *and* the outcomes obtained by them. To my knowledge, no formal model of imitation, adopting the re-

[11] A public good is a collective good that requires individual contributions to be produced. However, the individuals who do not contribute to its creation cannot be excluded from its consumption. The production of a public good is typically modeled as an interaction resembling a prisoner's dilemma with more than two players, where for every player not contributing to the public good (i.e., free riding) is a dominant strategy. A general introduction to this literature can be found in Fehr and Gintis (2007), and Gächter (2008).

strictive definition of imitation proposed by Barrera and Buskens [2007], is available. However, such a model would be useful not only for the study of trust in embedded settings, but also for other social phenomena, for example conformist behavior. The advantage of such an imitation model is that the underlying behavioral theory is quite simple and the model would not be very demanding in terms of cognitive abilities of the actors.

When the student in my example drives into the parking-lot of the university, she may decide to trust the attendant just because she can easily see that many other students left their cars to the attendant. In spite of its plausibility, however, this imitation mechanism can have perverse effects, since the student is not likely to find out the day after whether any of those cars was stolen or damaged. An example of an imitation mechanism leading to dramatically perverse effect can be found in Franzinelli's book about betrayal in Italy under the fascist regime. In 1938, after the Italian government enacted the "leggi razziali" (racial laws), Italian Jews started to be discriminated and – later – deported. Franzinelli [2001, 178-179] describes that many Jews tried to flee from the country. Consequently, smugglers started to organize their escape. Some smugglers were indeed trying to help the Jews, but others aimed at taking possession of their property and handed the Jews over to the Germans. Given the high risk of staying in Italy, many Jews had to trust the smugglers in order to be guided to the Swiss border. Often, these Jews obtained information about the smugglers through informal networks, such as relatives or friends of other Jews who had already left, but they could hardly obtain reliable information on whether these other Jews got across the border safely. In particular, they did not know how many of them were ultimately handed over to the Nazis.

4. Empirical Research on the Social Mechanisms of Trust

In the last two decades, empirical research on trust in embedded settings has been abundant [see Buskens and Raub 2008 for a detailed review]. Moreover, various research methods have been applied, including surveys [e.g. Gulati 1995], laboratory experiments [e.g. Bolton *et al.* 2004], factorial surveys [e.g. Rooks *et al.* 2000], and combinations of complementary methods [e.g. Simpson and McGrimmon 2008]. Different research methods have different assets and liabilities. Laboratory experiments are, by definition, more efficient at capturing *actual* behaviour as they allow observation of the actors' decisions in perfectly controlled environment [see Willer and Walker 2007, ch. 1, on the role of experimentation in sociology]. By contrast, surveys maximize the possibility to capture the initial conditions in which the actions

take place but provide information only on personal attitudes rather than on behaviour. Since this paper focuses on the *Social Mechanisms* of trust, in this section I will concentrate my discussion only on laboratory experiments on embedded trust, because in general, in surveys and factorial surveys, the actual mechanism is explicitly or implicitly assumed rather than empirically observed.

4.1. *Dyadic Embeddedness*

Experiments studying dyadic embeddedness typically apply finitely or indefinitely repeated games.[12] Camerer and Weigelt [1988], Neral and Ochs [1992], Anderhub *et al.* [2002], and Brandt and Figueras [2003] ran experiments using a finitely repeated trust game. Effects of *dyadic control* based on Kreps and Wilson [1982] model with incomplete information were found consistently in all these experiments: trust is generally placed and honored in the first rounds of the repeated game, then both trust and trustworthiness collapse when the last rounds approach. Effects of *dyadic learning* are generally supported, too: actors do not place trust when it has been abused in the previous round. These experiments reproduce the model's predictions pretty well although there are some deviations [see Camerer 2003, 446-453, for a detailed discussion of these experiments]. Similar results were found also by Gautschi [2000] who used a shorter series of repeated Trust Games and by Kollock [1994] who likewise ran shorter series but using a different experimental design, where the interaction was framed as a transaction between a buyer and a seller. The *indefinitely* repeated Trust Game was studied by Engle-Warnick and Slonim [2004] who found that trust decreases over time, although in theory this should not happen if players do not know when the end of the repeated interaction is coming. One explanation for this anomalous result might be that as the game proceeds the players believe that the probability that it will end increases, even though they are informed that it does not.[13] Finally, Barrera [2007] found an effect of dyadic learning in a one-shot Investment Game, comparing pairs with and without a common past, where the common past was created by letting the actors play a bargaining game before the Investment Game. However, this effect of learning was significant only for trustors and not for trustees.

[12] The literature on one-shot games is extremely vast, but I do not discuss these experiments here. An overview of these studies can be found in Camerer 2003, 83-100.

[13] This misconception is equivalent to the popular belief that "late" numbers – numbers who have not been extracted for longer time at the Italian Lotto – are more likely to be extracted, even though everybody should know that the probability is always the same since extractions are independent.

4.2. *Network Embeddedness*

Investigating the effects of network embeddedness on trust using experiments is a rather complex enterprise, because implementing personal relationships in the laboratory is virtually impossible. Accordingly, the few studies who investigated this problem manipulated social networks in terms of information transmission: the researchers introduced the network by letting the computer program used in the experiments transmit the information among the players. For example, in a network of trustors, every trustor can see on her screen what choices were made by the other trustors with whom she shares a network tie.

Bolton *et al.* [2004] ran a study with three treatments, which they call, stranger, partner, and reputation. In the stranger treatment, the participants played series of one-shot Trust Games, and where matched with a different player in every round. In the partner treatment, the participants played a finitely repeated Trust Game, always with the same partner. In the reputation treatment, the participants played series of one-shot Trust Games, like in the stranger treatment, but every player received information on the choices made by all other players in previous rounds. Thus, the reputation treatment is equivalent to a fully connected network of information transmission. Bolton *et al.* [2004] found that both trust and trustworthiness collapse rather soon in the stranger treatment. The partner treatment displays the typical results of finitely repeated Trust Games, with high trust and trustworthiness in the first rounds, followed by no trust or abuse of trust towards the last rounds. Finally, in the reputation treatment, trust starts slower, but builds up as the game proceeds until it stabilizes at a somewhat lower level than in the partner treatment. This indicates that the players realize that they have more to gain if they keep placing and honoring trust. Bohnet and Huck [2004] used a similar experimental design and obtained similar results.

However, although these studies clearly show that network embeddedness promotes trust, the mechanism responsible for the positive effect is not pinpointed. On the one hand, trustees may honor trust because they care about their reputation and trustors may anticipate on this and place trust accordingly. On the other hand, the trustors may learn from the reputation score of their partner to what extent he is trustworthy and place trust accordingly. In other words, these results are consistent with both learning and control. An experiment in which all mechanisms were put simultaneously to a test was run by Barrera and Buskens [2009]. In their study, Barrera and Buskens let groups of six subjects – four trustors and two trustees – play a finitely repeated Investment Game in which the trustors received information about the behavior of the other players. Furthermore, they manipulated the information transmitted along these small networks so that some trustors received information on

the behavior of both another trustor and her partner, while other trustors received information only on the behavior of another trustor, but not on the behavior of the trustee interacting with this trustor. This experimental manipulation permitted to test simultaneously all mechanisms as well as to disentangle effects of learning and imitation.

Next to dyadic learning and dyadic control, Barrera and Buskens found empirical evidence for network learning and imitation, but no evidence for network control. Furthermore, surprisingly, the effect of imitation was stronger than the effect of learning. Interpreting these results, it seems that as soon as the complexity of the interaction and the amount of information that the players need to process increase, the actors begin to adopt cognitively simpler heuristics. Thus, the effect of learning is stronger at the beginning and becomes weaker in later series of games while the effect of imitation shows the opposite pattern.

5. Conclusion

In this paper, I summarized the literature on the social mechanisms of trust. In the first part of the paper I discussed the link between the analytical approach to trust and a number of prominent definitions of trust found in the literature [e.g., Coleman 1990, ch. 5; Gambetta 1988; Hardin 2002]. Subsequently, I reviewed some actor-based theoretical models of trust in embedded settings which lead to identify two main social mechanisms, *learning* and *control* [Yamagishi and Yamagishi 1994; Buskens 2002; Buskens and Raub 2002]. In addition, I discussed a third mechanism, *imitation* [Barrera 2005; Barrera and Buskens 2007], for which some empirical evidence is available [Barrera and Buskens 2008], but no formal model have yet been developed. Finally, in the last part of the paper, I briefly summarized the experimental research on these social mechanisms.

In my view, there are at least three advantages resulting from the analytical approach to the study of trust. First, the analytical approach is intrinsically actor-based. Therefore, it is best suitable for designing policies which ideally should likewise be actor-based. Since the importance of trust as a "lubricant for cooperation" [Arrow 1974] is widely recognized, understanding which mechanisms drive the development of trust is crucially important for organizations as well as for the society at large, because both have a strong interest in promoting cooperation between their members. In general policies are more likely to be effective if they target the individual actors and operate on the right leverages.

Second, understanding social mechanisms implies opening the black box of causal processes affecting social phenomena. Such causal processes cannot be appreciated if social research is conducted mainly at the aggregated level; because different configurations of actors and interactions can potentially lead to superficially similar collective outcomes [see Hedström 2005 on this point].

Third, the social mechanism approach emphasizes the importance of the micro level, by requiring that the behavioral theory is made explicit in order to give a complete (causal) account of a social phenomenon. The identification of the main assumptions for a general behavioral theory is very important for the process of unification of the discipline around a common paradigm [cf. Boudon 2002]. In this paper, I make a contribution to this point by discussing the social mechanisms of trust in a stepwise fashion, underlining the key assumptions that distinguish the mechanisms from each other.

Finally, I want to sketch some possible direction for future research on the social mechanisms of trust. First, the three mechanisms analyzed in the paper focus primarily on the role of the trustor in trust problems. In general, the mechanism of control simply assumes that trustees anticipate on the decision of the trustors and behave accordingly, while learning and imitation are usually coupled with the assumption of incomplete information (i.e., the existence of different "types" of trustees: trustworthy and untrustworthy ones). However, since the trustor's decision to place trust is generally assumed to depend on her assessment of the trustee's trustworthiness [Gambetta 1988; Coleman 1990, ch. 5], it seems particularly important to understand what mechanisms drive the actions of the trustees. Nevertheless, research on the determinants of the behavior of the trustees is still rather scarse [e.g., Buskens, Raub, and van der Veer 2008 and Barrera 2007].

Second, although evidence of imitation was found in some empirical studies [Barrera and Buskens 2007; 2009], from a theoretical point of view imitation has been taken as a simple heuristic rather than explicitly modeled. Furthermore, although imitation is certainly very common, a systematic investigation of the social conditions under which imitation is more likely to occur is still lacking.

Third, although the importance of social networks for trust problems and for the solution of cooperation problems in general is now widely recognized, all the literature discussed in this paper treats social networks as exogenous. If the advantages of network embeddedness are perceived also by the actors themselves, then we should expect that the actors actively invest in the creation of social networks [Coleman 1990, ch. 12; Flap 2004; Buskens and van de Rijt 2008]. In fact, research on interactions in embedded settings in which networks are endogenized is rather recent, but growing quite rapidly [e.g. Goyal 2007].

References

Altmann, M.

1993 "Reinterpreting Network Measures for Models of Disease Transmission." *Social Networks* 15: 1-7.

Anderhub, V., Engelmann D., and Güth, W.

2002 "An Experimental Study of the Repeated Trust Game with Incomplete Information." *Journal of Economic Behavior and Organization* 48: 197-216.

Arrow, K.J.

1974 *The Limits of Organizations*. New York: Norton.

Axelrod, R.

1984 *The Evolution of Cooperation*. New York: Basic Books.

Barber, B.

1983 *The Logic and Limits of Trust*. New Brunswick, NJ: Rutgers University Press.

Barbera, F.

2004 *Meccanismi Sociali. Elementi di sociologia analitica*. Bologna: Il Mulino.

Barrera, D.

2005 *Trust in Embedded Settings*. Veenendaal: Universal Press.

2007 "The Impact of Negotiated Exchange on Trust and Trustworthiness." *Social Networks* 29: 508-526.

Barrera, D., and Buskens, V.

2007 "Imitation and Learning under Uncertainty: A Vignette Experiment." *International Sociology* 22: 366-395.

2009 "Third-Party Effects in an Embedded Investment Game." Forthcoming in *Trust and Reputation*, edited by V. Buskens, C. Cheshire, K.S. Cook, and C. Snijders. New York: Russell Sage Foundation.

Bandura, A., and Walters, R.H.

1963 *Social Learning and Personality Development*. New York: Holt, Reinehart and Winston Inc.

Berg, J.E., Dickhaut, J., and McCabe, K.

1995 "Trust, Reciprocity, and Social History." G*ames and Economic Behavior* 10: 122-142.

Bijlsma, K., and Costa, A.C.

2005 "Understanding the Trust-control Nexus." *International Sociology* 20: 259-282.

Bijlsma, K.M., and Van de Bunt, G.G.

2003 "Antecedents of Trust in Managers: A "Bottom-Up" Approach." *Personnel Review* 32: 638-664.

Binmore, K.

1998 *Game Theory and the Social Contract, Volume 2: Just Playing*. Cambridge, MA: MIT Press.

Bohnet, I., and Huck, S.

2004 "Repetition and Reputation: Implications for Trust and Trustworthiness When Institutions Change." *American Economic Review (Papers and Proceedings)* 94: 362-366.

Bolton, G.E., and Ockenfels, A.

2000 "ERC: A Theory of Equity, Reciprocity, and Competition." *American Economic Review* 90: 166-193.

Bolton, G.E., Katok, E., and Ockenfels, A.

2004 "How Effective Are Electronic Reputation Mechanisms? An Experimental Investigation." *Management Science* 50: 1587-1602.

Boudon, R.

1986 *Theories of Social Change: A Critical Appraisal*. Cambridge: Polity Press.

2002 "Sociology that Really Matters." *European Sociological Review* 18: 371-378.

Bower, A., Garber, S., and Watson, J.C.

1997 "Learning about a Population of Agents and the Evolution of Trust and Cooperation." *International Journal of Industrial Organization* 15: 165-190.

Brandts, J., and Figueras, N.

2003 "An Exploration of Reputation Formation in Experimental Games." *Journal of Economic Behavior and Organization* 50: 89-115.

Burt, R.S.

2005 *Brokerage and Closure. An Introduction to Social Capital*. New York: Oxford University Press.

Burt, R.S., and Knez, M.

1995 "Kinds of Third-Parties Effects on Trust." *Rationality and Society* 7: 255-292.

Buskens, V.

2002 *Trust and Social Networks*. Boston: Kluwer.

Buskens, V., and Raub, W.

2002 "Embedded Trust: Control and Learning." *Advances in Group Processes* 19: 167-202.

2008 "Rational Choice Research on Social Dilemmas." Forthcoming in *Handbook of Rational Choice Social Research*, edited by R.P.M Wittek, T.A.B. Snijders and V. Nee. New York: Russell Sage Foundation.

Buskens, V., Raub, W., and van der Veer, J.

2008 "Trust in Triads: An Experimental Study." Working paper.

Buskens, V. and van de Rijt, A.

2008 "Dynamics of Networks if Everyone Strives for Structural Holes". Forthcoming in .

Buskens, V., and Weesie, J.

2000 "Cooperation via Networks." *Analyse und Kritik* 22: 44-74.

Camerer, C.F.

2003 *Behavioral Game Theory*. Princeton, NJ: Princeton University Press.

Camerer, C.F., and Weigelt, K.

1988 "Experimental Tests of a Sequential Equilibrium Reputation Model." *Econometrica* 56: 1-36.

Coleman, J.S.

1990 *Foundation of Social Theory*. Cambridge, MA: The Belknap Press of Harvard University Press.

Craswell, R.

1993 "On the Uses of 'Trust': Comment on Williamson, 'Calculativeness, Trust, and Economic Organization.' " *Journal of Law and Economics* 36: 487-500.

Das T.K., and Teng, B.

1998 "Between Trust and Control: Developing Confidence in Partner Cooperation in Alliances." *Academy of Management Review* 23: 491-512.

Dasgupta, P.

1988 "Trust as a Commodity. " Pp. 49-72 in *Trust: Making and Breaking Cooperative Relations*, edited by D. Gambetta. Oxford: Blackwell.

Engle-Warnick, J., and Slonim, R.L.

2004 "The Evolution of Strategies in a Repeated Trust Game." *Journal of Economic Behavior and Organization* 55: 553-573.

Fehr, E., and Gintis, H.

2007 "Human Motivation and Social Cooperation: Experimental and Analytical Foundations." *Annual Review of Sociology* 33: 43-64.

Fehr E., and Schmidt, K.M.

1999 "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114: 817-868.

Flap, H.

2004 "Creation and Returns of Social Capital. " Pp. 3-23 in *Creation and Returns of Social Capital*, edited by H. Flap and B. Völker. London: Routledge.

Franzinelli, M.

2001 *Delatori. Spie e confidenti anonimi: l'arma segreta del regime fascista*. Milano: Mondadori.

Gächter, S.

2008 "Rationality, Social Preferences and Strategic Decision-Making from a Behavioral Economics Perspective." Forthcoming in *Handbook of Rational Choice Social Research*, edited by R.P.M Wittek, T.A.B. Snijders and V. Nee. New York: Russell Sage Foundation.

Gambetta, D.

1988 "Can We Trust Trust?" Pp. 213-237 in *Trust: Making and Breaking Cooperative Relations*, edited by D. Gambetta. New York: Blackwell.

Gautschi, T.

2000 "History Effects in Social Dilemma Situations." *Rationality and Society* 12: 131-162.

Goyal, S.

2007 *Connections. An Introduction to the Economics of Networks*. Princeton, NJ: Princeton University Press.

Granovetter, M.

1985 "Economic Action and Social Structure: The Problem of Embeddedness." *American Journal of Sociology* 91: 481-510.

Gulati, R.

1995 "Does Familiarity Breed Trust? The Implications of Repeated Ties for Contractual Choice in Alliances." *Academy of Management Journal* 38: 85-112.

Hardin, R.

2002 *Trust and Trustworthiness*. New York: Russell Sage Foundation.

Harsanyi, J.C.

1967-68 "Games with Incomplete Information Played by 'Bayesian' Players I-III." *Management Science* 14:159-182, 320-334, 486-502.

Hedström, P.

1998 "Rational Imitation." Pp. 306-327 in *Social Mechanism: An Analytical Approach to Social Theory*, edited by P. Hedström, and R. Swedberg. Cambridge: Cambridge University Press.

2005 *Dissecting the Social. On the Principles of Analytical Sociology*. Cambridge: Cambridge University Press.

Hedström, P., and Swedberg, R. (eds.)

1998 *Social Mechanism: An Analytical Approach to Social Theory*. Cambridge: Cambridge University Press.

Hirschman, A.O.

1970 *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Cambridge, MA: Harvard University Press.

Kollock, P.

1994 "The Emergence of Exchange Structures: An Experimental Study of Uncertainty, Commitment, and Trust." *American Journal of Sociology* 100: 313-345.

Kreps, D.M.

1990 "Corporate Culture and Economic Theory." Pp. 90-143 in *Perspective on Positive Political Economy*, edited by J. Alt and K. Shepsle. Cambridge: Cambridge University Press.

Kreps, D.M., and Wilson, R.

1982 "Sequential Equilibria." *Econometrica* 50: 863-894.

Lahno, B.

1995 "Trust, Reputation, and Exit in Exchange Relationships." *Journal of Conflict Resolution* 39: 495-510.

Lewis, J.D., and Weigert, A.

1985 "Trust as a Social Reality." *Social Forces* 63: 967-985.

Luhmann, N.

1988 "Familiarity, Confidence, Trust: Problems and Alternatives." Pp. 94-107 in *Trust: Making and Breaking Cooperative Relations*, edited by D. Gambetta. Oxford: Blackwell.

Macy, M.W., and Flache, A.

1995 "Beyond Rationality in Models of Choice." *Annual Review of Sociology* 21: 73-91.

2002 "Learning Dynamics in Social Dilemmas." *Proceedings of the National Academy of Science* 99: 7229-7236.

Macy, M.W., and Skvoretz, J.

1998 "The Evolution of Trust and Cooperation between Strangers: A Computational Model." *American Sociological Review* 63: 638-660.

Misztal, B.A.

1996 *Trust in Modern Societies*. Cambridge: Polity Press.

Möllering, G.

2005 "The Trust/Control Duality. An Integrative Perspective on Positive Expectations of Others." *International Sociology* 20: 283-305.

Neral, J., and Ochs, J.

1992 "The Sequential Equilibrium Theory of Reputation Building: A Further Test." *Econometrica* 60: 1151-1169.

Nooteboom, B.

2002 *Trust: Forms, Foundations, Functions, Failures, and Figures*. Northampton, MA: Edward Elgar.

Ortmann, A., Fitzgerald, J., and Boeing, C.

2000 "Trust, Reciprocity, and Social History: A Re-Examination." *Experimental Economics* 3: 81-100.

Parsons, T.

1937 *The Structure of Social Action*. New York: McGraw-Hill.

Pingle, M.

1995 "Imitation versus Rationality: An Experimental Perspective on Decision Making." *Journal of Socio-Economics* 24: 281-316.

Pingle, M., and Day, R.H.

1996 "Modes of Economizing Behavior: Experimental Evidence." *Journal of Economic Behavior and Organization* 29: 191-209.

Rabin, M.

1993 "Incorporating Fairness into Game Theory and Economics." *American Economic Review* 83: 1281-1302.

Rapoport, A.

1974 "Prisoner's Dilemma – Recollections and Observation." Pp. 17-34 in *Game Theory as a Theory of Conflict Resolution*, edited by A. Rapoport. Dordrecht: Reidel.

Rasmusen, E.

2001 *Games and Information: An Introduction to Game Theory*. Oxford: Blackwell, Third edition.

Raub, W.

1997 *Samenwerking in Duurzame Relaties en Sociale Cohesie*. Amsterdam: Thesis Publishers.

2004 "Hostage Posting as a Mechanism of Trust: Binding, Compensation, and Signaling." *Rationality and Society* 16: 319-366.

Raub, W., and Weesie, J.

1990 "Reputation and Efficiency in Social Interactions: An Example of Network Effects." *American Journal of Sociology* 96: 626-654.

2000 "The Management of Matches: A Research Program in Durable Social Relations." *The Netherlands Journal of Social Sciences* 36: 71-88.

Robinson, S.L.

1996 "Trust and Breach of the Psychological Contract." *Administrative Science Quarterly* 41: 574-599.

Rooks, G., Raub, W., Selten, R., and Tazelaar, F.

2000 "Cooperation between Buyer and Supplier: Effects of Social Embeddedness on Negotiation Effort." *Acta Sociologica* 43: 123-137.

Rousseau, M.T., Stikin, S.B., Burt, S.B., and Camerer, C.

1998 "Not So Different After All: Across-Discipline View of Trust." *Academy of Management Review* 23: 393-404.

Rus, A., and Iglič, H.

2005 "Trust, Governance and Performance." *International Sociology* 20: 371-391.

Schlag, K.

1998 "Why Imitate, and If So, How? A Bounded Rational Approach to Multi-Armed Bandits." *Journal of Economic Theory* 78: 130-156.

Simpson, B., and McGrimmon, T.

2008 "Trust in Embedded Markets: A Multi-method Investigation of Consumer Transactions." *Social Networks* 30: 1-15.

Snijders, C.

1996 *Trust and Commitments*. Amsterdam: Thela Thesis.

Snijders, C., and Keren, G.

2001 "Do You Trust? Whom Do You Trust? When Do You Trust?" *Advances in Group Processes* 18: 129-160.

Valente, T.W.

1995 *Network Models of the Diffusion of Innovations*. Cresskill: Hampton Press.

Van de Bunt, G.G., Wittek, R.P.M., and de Klepper, M.C.

2005 "The Evolution of Intra-Organizational Trust Networks: The Case of a German Paper Factory: An Empirical Test of Six Trust Mechanisms." *International Sociology* 20: 339-370.

Weesie, J.

1996 "Disciplining via Exit and Voice." ISCORE paper n. 88. Utrecht University.

Willer, D., and Walker, H.A.

2007 *Building Experiments. Testing Social Theories*. Stanford, CA: Stanford University Press.

Williamson, O.E.

1993 "Calculativeness, Trust, and Economic Organization." *Journal of Law and Economics* 36: 453-486.

Wittek, R.P.M.

2001 "Mimetic Trust and Intra-Organizational Network Dynamics." *Journal of Mathematical Sociology* 25: 109-138.

Yamagishi, T., and Yamagishi, M.

1994 "Trust and Commitment in the United States and Japan." *Motivation and Emotion* 18: 129-166.

The Social Mechanisms of Trust

Abstract: In the last decades, problems of trust and cooperation in general have received much attention from scholars working in various scientific disciplines. In particular, research in the field of analytical sociology has focused on the emergence of trust in embedded settings investigating the individual decisions of the actors involved. These studies have lead to the identification of three social mechanisms affecting trust in embedded settings: Control, learning, and imitation. In this paper, I review the main theoretical models underlying these mechanisms, discuss the link between these models and a number of prominent definitions of trust found in the literature, and review the experimental research on these mechanisms.

Keywords: social mechanisms, trust, game theory, social networks.

Davide Barrera is Assistant Professor at the Department of Sociology/ICS, Utrecht University, where he received his PhD in 2005. His ares of research interests include social mechanisms, behavioral game theory, cooperation problems, and social networks. His work is generally interdisciplinary, and applies various research techniques like surveys, vignette experiments, and, especially laboratory experiments.